

*Citation for published version:*

Tee, P, Wakeman, I, Parisi, G, Dawes, J & Kiss, IZ 2017, 'Constraints and entropy in a model of network evolution', *European Physical Journal B : Condensed Matter and Complex Systems*, vol. 90, no. 11, 226, pp. 1 - 16. <https://doi.org/10.1140/epjb/e2017-80185-5>

*DOI:*

[10.1140/epjb/e2017-80185-5](https://doi.org/10.1140/epjb/e2017-80185-5)

*Publication date:*

2017

*Document Version*

Peer reviewed version

[Link to publication](#)

*Publisher Rights*

CC BY

The final publication is available at Springer via: <https://doi.org/10.1140/epjb/e2017-80185-5>

**University of Bath**

**Alternative formats**

If you require this document in an alternative format, please contact:  
[openaccess@bath.ac.uk](mailto:openaccess@bath.ac.uk)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Constraints and Entropy in a Model of Network Evolution

Philip Tee<sup>1,2</sup>, Ian Wakeman<sup>2</sup>, George Parisi<sup>2</sup>, Jonathan Dawes<sup>3</sup>, and István Z. Kiss<sup>4</sup>

<sup>1</sup> Moogsoft Inc,

1265 Battery Street, San Francisco, California 94111 USA e-mail: phil@moogsoft.com

<sup>2</sup> School of Engineering and Informatics,

University of Sussex, Brighton BN1 9RH UK e-mail: ianw,g.parisi,p.tee@sussex.ac.uk

<sup>3</sup> Centre for Networks and Collective Behaviour, and

Department of Mathematical Sciences,

University of Bath, Bath BA2 7AY, UK e-mail: j.h.p.dawes@bath.ac.uk

<sup>4</sup> Department of Mathematics, School of Mathematical and Physical Sciences,

University of Sussex, Brighton BN1 9QH, UK e-mail: i.z.kiss@sussex.ac.uk

the date of receipt and acceptance should be inserted later

**Abstract.** Barabási-Albert’s ‘Scale Free’ model is the starting point for much of the accepted theory of the evolution of real world communication networks. Careful comparison of the theory with a wide range of real world networks, however, indicates that the model is in some cases, only a rough approximation to the dynamical evolution of real networks. In particular, the exponent  $\gamma$  of the power law distribution of degree is predicted by the model to be exactly 3, whereas in a number of real world networks it has values between 1.2 and 2.9. In addition, the degree distributions of real networks exhibit cut offs at high node degree, which indicates the existence of maximal node degrees for these networks. In this paper we propose a simple extension to the ‘Scale Free’ model, which offers better agreement with the experimental data. This improvement is satisfying, but the model still does not explain *why* the attachment probabilities should favor high degree nodes, or indeed how constraints arrive in non-physical networks. Using recent advances in the analysis of the entropy of graphs at the node level we propose a first principles derivation for the ‘Scale Free’ and ‘constraints’ model from thermodynamic principles, and demonstrate that both preferential attachment and constraints could arise as a natural consequence of the second law of thermodynamics.

## 1 Introduction and Background

### 1.1 Overview

The ‘*Scale Free*’ model of Barabási-Albert [1] is widely accepted as the definitive model of how real world networks evolve. This and other dynamic network models consider real world networks as graphs  $G(V, E)$ , where  $V(t)$  is the set of vertices and  $E(t)$  the set of edges. Its success at overcoming the difficulties of applying the Erdős-Rényi (ER) random graph model (for a detailed description see [2]) to real world networks is well understood. In particular the model naturally results in a power law degree distribution, as opposed to the random graph model, which has a binomial distribution of node degree, which in the continuum limit of a very large network is approximately Poisson, with well defined higher statistical moments that establish the ‘scale’ of the graph. This is in stark contrast to the scale free model which does not have well defined moments above the mean. The model described by Barabási-Albert in [3] and [1] builds upon, and provides an explanation for, the notion of the small world network, first introduced by Watts and Strogatz [4] and has been used to analyze a wide variety of real world graphs.

On close examination, the scale free model has a number of theoretical challenges, and, it is well understood that the behavior of real world networks has deeper complexity than a single constant power law degree distribution. Of course balanced against the success of the model in generating networks that share the *small world* property and scale free degree distributions, these challenges can be viewed as opportunities for refinement of the fundamental approach. In this work we focus on extensions to the model which provide improvements in the following three areas:

- *Absence of Constraints*: There is an assumption that a graph can continue to evolve indefinitely, unconstrained by any system wide or external resources. For most real world networks this is not the case. For example in communication networks every node in the network has a natural maximum connectivity. In the scale free model there is no such upper limit to node degree.
- *Fit to Real World Data*: The standard scale free model produces a degree distribution that follows a power law with exponent  $\gamma = 3$ . It is well understood that this is not an exact fit to real world data, which we highlight

in Section 3. Many extensions exist that produce a better fit, some of which we survey later. It is clear that the degree distributions of real networks have more complex behavior than a simple fixed exponent power law.

- *Absence of a Physical Model:* The notion of scale freedom derives directly from the hypothesis of preferential attachment, that is in a dynamically evolving graph new nodes will more likely attach to nodes of higher degree. Whilst the scale free model provides a theoretical framework that points to high node degree making a node more likely to attract new connections, there is no fundamental explanation of *why* that should be so, and what physical processes may be at work that could produce that effect. It would be desirable if this could be explained using a first principles argument involving well understood mechanisms. This would further strengthen the fundamental premise of the scale free model.

In this paper we will attempt to address these challenges. We do so by proposing a simple extension to the standard scale free model, which introduces a hard cut off in the degree of a node, motivated by considerations from communications network design. This model has some attractive features, amongst which is a more accurate prediction of the power law exponent. Although extensions to the preferential attachment approach (most notably [5], [6] and [7]), can result in values of the power law exponent less than 3, we believe our model achieves this through a simple and natural extension to the traditional preferential attachment paradigm. Furthermore, as a consequence of introducing the constraint, we identify that the attachment probability introduces *superlinear* polynomial terms in node degree. This additional structure to the attachment probability is responsible for a richer scaling regime in node degree evolution. This structure allows us to compare in Section 4 both the constraints and scale free model to a novel model of evolution that argues from a stochastic perspective based upon recent developments in the structural entropy of a graph. By developing the outline of an entropic model we illustrate how both the standard scale free and our constrained model could be viewed as approximations to a more fundamental, statistical thermodynamic model of network growth.

In this section we will begin with a brief overview of the continuum analysis used in [1] to derive the principle results of scale free models, and at a very high level subsequent attempts to build upon and extend the model. We will make use of the same continuum approximation in our analysis.

We show in section 2 how the introduction of a simple environmental constraint into the scale free model can significantly improve its predictive power, and compare our *constrained* model to a range of more contemporary network data in section 3. As part of the verification of our *constrained* model, we also present results of simulations of network growth using our modified attachment probability defined in Section 2. An attractive feature of our extended model is that it reproduces the scale free model

when we allow our constraint to tend to infinity. We are able to significantly outperform the ability of the scale free model to predict the exponent  $\gamma$  of the power law distribution across a wide range of real world data (results are summarized in Table 2). In particular for ten of the twenty three data sets analyzed (marked in Table 2 in bold) we are able to predict  $\gamma$  to within 10%, whereas the scale free model overestimates the value of  $\gamma$  by an average of 35% and in only four cases does it predict within the range 10-20%. Our constrained model therefore performs better than the standard scale free model on the first two issues identified above, but not on the third.

In Section 4 we propose a novel statistical thermodynamical (i.e. entropic) model of network growth. This addresses the third objective. Recent work on the behavior of communications networks by Tee *et al* [8,9] introduced a measure of the structural entropy of a node, derived from its degree and clustering coefficient. We show how this can lead to a direct derivation of scale free and constraint models, potentially explaining why scale freedom arises and why our constrained model is a better fit for networks as they grow and encounter connectivity limitations. We present in the same section some early results from numerical simulations of the entropic model, which show many of the features of the real world data we analyzed in Section 3.

## 1.2 The Scale Free Model

The Scale Free Model of Barabási, Albert and Jeong [3], [1] is based on two simple and fundamental assumptions:

- *Growth:* Starting with  $m_0$  nodes and  $e_0$  edges, we add a new node at each unit time step. When this node is added to the network, it connects to  $m \ll m_0$  other nodes. This process continues indefinitely, such that after  $t$  unit time steps, there are  $m_0 + t$  nodes, and  $e_0 + mt$  edges. Eventually the constants in these expressions can be dropped as they are insignificant compared to  $t$ .
- *Preferential Attachment:* The node attaches to other nodes with a probability determined by the degree of the target node, such that more highly connected nodes are *preferred* over lower degree nodes.

Using a mean field theory approach the analysis explains both the power law scaling of real world networks [10], and the simultaneous resilience and vulnerability of networks to random and targeted attacks, respectively [11]. The approach taken in [3] begins by proposing the probability of a *randomly chosen node*  $i$ , capturing a connection to a new node, as solely dependent upon its degree  $k_i$  as:

$$\Pi_i = \frac{k_i}{\sum_j k_j} = \frac{k_i}{2mt}, \quad (1)$$

In the strictest sense the approximation  $\sum_j k_j = 2mt$  should include the original nodes  $m_0$  and their degrees, however for large values of  $t$  this can be effectively ignored, without loss of generality, as  $2e_0 \ll 2mt$ . By taking the continuous approximation, this naturally leads to

the following ordinary differential equation for the time evolution of node  $i$ 's degree  $k_i(t)$ :

$$\frac{dk_i(t)}{dt} = m\Pi_i = \frac{k_i(t)}{2t}. \quad (2)$$

Equation (2), can be solved subject to an initial condition that at time  $t_i$ , when node  $i$  is added, its degree  $k_i = m$  to yield:

$$k_i(t) = m \left( \frac{t}{t_i} \right)^{1/2}. \quad (3)$$

In order to derive the degree distribution begin by assuming that  $t$  is fixed. At this stage the probability that  $k_i(t)$  is smaller than a given degree  $k$  is:

$$P(k_i(t) < k) = P\left(t_i > \frac{m^2 t}{k^2}\right) = 1 - P\left(t_i \leq \frac{m^2 t}{k^2}\right).$$

Developing the mean field approach we note that the  $i$ th node was chosen at random, so its time of introduction into the network  $t_i$  is a random variable. Given that nodes are added at each time step, the range of possible values for  $t_i$  are  $1, 2, \dots, (m_0 + t)$ , and each value can occur with probability  $\frac{1}{(m_0 + t)}$ . We can conclude that the random variable  $t_i$  is uniformly distributed and can write the probability of choosing a node  $i$  with a  $t_i$  smaller than  $\frac{m^2 t}{k^2}$  as:

$$P\left(t_i \leq \frac{m^2 t}{k^2}\right) = \frac{1}{(m_0 + t)} \times \frac{m^2 t}{k^2}.$$

We can now state that the probability of a node having degree  $k < k_i$  as:

$$P(k_i(t) < k) = 1 - \frac{m^2 t}{k^2(m_0 + t)} = \int_m^k P(x) dx,$$

implying that  $P(k) = \frac{\partial P(k_i(t) < k)}{\partial k}$ , yielding the principal result of the Barabási-Albert Scale Free model:

$$P(k) = \frac{2m^2 t}{m_0 + t} \frac{1}{k^3}. \quad (4)$$

This predicts that on a log/log scale the slope of the degree distribution  $\gamma$  is identically 3. The result has been compared against many real world networks, and indeed the power law behavior has been seen in many examples and is one of the triumphs of the scale free model. The model, however, generally overestimates the value of  $\gamma$  and cannot explain the non linear behavior of the degree distribution at high values of  $k$  (as outlined in [12]). Reproduced in Table 1 from the data in [1] are some key parameters from a selection of the analyzed real world networks. The data is taken from a wide range of sources, which we supplement in Section 3, including the classic movie actor collaboration network from IMDB, a physical communications network, a biological network and a number of collaboration

networks. A striking feature of all of these networks is both a limit to the degree of a node, and also that the value of  $\gamma$  is significantly lower than predicted by the scale free model ( $\gamma$  is calculated as described in Section 3.1.). Recent work [13] has highlighted a number of deficiencies in the scale free model, including deviations from the scale free degree distributions and the presence of cut offs in the maximum degree. It must be stated however that the model is strikingly powerful in its ability from a simple set of assumptions to explain many features of complex networks, from their small world property to the absence of a ‘scale’ in the degree distributions. This simplicity is powerful and hints at fundamental processes underlying the dynamics of network evolution.

Failure to capture the detail of the degree distributions of real world networks however, indicates that this simplicity must be supplemented with additional facets to the model of node attachment. In addition the appeal to node degree being the primary determinant of attachment probability is a modeling assumption and does not explain *why* that is the case. The principal argument is based on the concept of “the rich get richer”, which is an equivalent statement to equation (1). In our view this is not a ‘first principles argument’, based upon fundamental physics. Given the success of the model and widespread acceptance of its validity and application in many fields from genetics to network design, it would be satisfying to link the derivation of equation (1) to core principles of physics. In this paper we start by exploring a next degree of approximation to the model to identify how environmental influences such as the presence of a top constraint for node degree alter the form of equation (1). In the model we propose this yields polynomial terms in  $k$ , which we hypothesize may be part of a series of corrections to the attachment probability. Using arguments based upon applying ensemble statistical mechanics to the entropy of a network vertex, we then propose an entropic model which naturally produces the concept of preferential attachment and constraints, and hints at further structure to the form of attachment probability in equation (1).

Table 1: Degree Distribution Parameters of some Real Networks [1]

Source	$\langle k \rangle$	Max Degree	$\gamma$
IMDB Movie Actors	28.78/127.33 <sup>1</sup>	900	2.3
Internet Router	2.57	30	2.48
Metabolic, <i>E. coli</i>	7.4	110	2.2
Co-authors, SPIRES	173	1100	1.2
Co-authors, neuro	11.54	400	2.1
Co-authors, math	3.9	120	2.5

### 1.3 Extensions to the Scale Free Model

Before embarking on an investigation of our model, it is important to stress that many proposals to extend prefer-

ential attachment have been advanced. These alternative models to preferential attachment rely upon modifications to the probability of attachment beyond simple dependence on the degree of the node. The extensions range from ecologically inspired models such as the competition based approach of D'Souza in [15], to direct alterations of the form of equation (1) by introducing 'super-linear' terms in  $k$ , that is arbitrary powers of  $k$ . The model of Krapivsky *et al* [7], explicitly explores forms of attachment probability where the term in  $k$  is replaced by an exponential form  $k^\alpha$ , where the exponent  $\alpha$  can vary in the range  $0 < \alpha < \infty$ . By varying  $\alpha$  it is possible to and produce very different forms of the degree distribution. These range from stretched exponential degree distribution to a super-linear zone for  $\alpha > 2$  where one node captures a connection to all other nodes. In other work, notably Dorogovtsev *et al* [5], the concept of initial attractiveness of a node is introduced, which permits values of the power law exponent to vary and produces values of  $\gamma$  that are between  $2 < \gamma < 3$ . These models depend upon the concept of some nodes starting with a higher initial attractiveness than others in their ability to gain connections to new nodes. In some ways this is the opposite approach to the constrained model we propose in this paper, where nodes become progressively less attractive as they acquire connections and approach their limit.

It is perhaps the ecological, and physically inspired extensions that are most attractive alternatives to preferential attachment. We have already mentioned the competition based model of D'Souza [15] that uses an optimization approach in which the minimization of a cost function upon every node addition is used to determine which node the new node attaches to. This model produces an exponentially corrected degree distribution of the form  $P(k) \propto k^{-\gamma} e^{-\alpha k}$ . This degree distribution is similar to that which we see in the data analyzed in Section 3, and is an encouraging advance on the original preferential attachment model.

Another widely accepted approach, which builds upon the work of Dorogovtsev, was developed by Barabási in collaboration with Bianconi. This model parametrizes the attractiveness of the node using a *fitness* measure,  $\eta_i$ , and was introduced in [6], [16] and further developed in the work of Moriano *et al* [17], and Su *et al* [18].

The extended model proposes that the probability of attachment is modified to include the fitness parameter in the most general sense, as follows:

$$\Pi_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j} . \quad (5)$$

To prevent this model requiring as many independent variables as there are nodes, the attractiveness  $\eta$  is fixed, or quenched, at node addition and is randomly assigned from an assumed probability distribution  $\rho(\eta)$  for the parameter. The model permits an analogy between the graph and the Bose-Einstein treatment of ideal gases. This analogy relies upon the identification of a node vertex with an energy level of the gas  $\epsilon_i$ , with the degree corresponding to the occupancy number of the energy level. Derivation

of graph properties from statistical mechanical arguments is long established, including in the work of Newman and Park on exponential random graphs described in [19]. In the Bianconi-Barabási model the fitness parameter is defined as  $\epsilon_i = -\frac{1}{\beta} \log \eta_i$ , with  $\beta$  being identified as classical inverse thermodynamic temperature. The denominator of equation (5) is then easily identified with the partition function  $Z$ , familiar from the Bose-Einstein model of statistical mechanics. Using the probability distribution  $\rho(\eta)$  of the nodes' fitness parameter as outline in [6],  $P(k)$  can be analytically solved for in the case of the uniform distribution to yield:

$$P(k) \sim \frac{k^{-1+C}}{\log(k)}, \text{ where } C \text{ is a constant} \quad (6)$$

This model is attractive, and indeed does provide a closer fit to the data, including the presence of a cut-off on the maximum degree of a node.

The models described thus far all share a similar set up to the original preferential attachment mechanism, in that they consider a stepwise addition of a single node which connects to a variable number of pre-existing nodes. In recent work by Bianconi *et al*, this has been generalized to investigate models based upon the addition of simplicial complexes to a network rather than nodes as described in [20,21]. These models, referred to as Network Geometry with Flavor (NGF), introduce the concept of a  $d$  dimensional simplex, which is a fully connected clique of  $d + 1$  nodes. When  $d = 1$  the model reduces down to the Bianconi-Barabási model, but higher dimensional simplices are hypothesized to more correctly represent the growth of networks where the unit of addition is a clique, such as a citation network being built from sub networks of frequently collaborating authors. The NGF model proceeds by adding a single node and links, so as to produce a new  $d$  dimensional simplex in the graph, by attaching the simplex to a randomly chosen  $d - 1$  existing face in the graph, governed by a generalized form of equation (5). The attachment probability is further parameterized by a flavor variable  $s$  which can take the values of  $-1, 0, 1$  that allows the introduction of a generalized degree which counts the number of  $d$  dimensional simplices incident to a node. The range of flavor ensures that the form of attachment probability, which is beyond the scope of this survey to outline, produces a well behaved probability. The survey in [20] has a full and complete overview of the model. The attraction of these models is the generation of a rich set of possible graph geometries, including scale free, Apollonian and a form of graph deeply analogous to the form of graphs proposed in a range of approaches to Quantum Gravity.

Together with the competition model of D'Souza these more physically and ecologically inspired models provide motivation to explore other analogies with such processes to improve upon the standard preferential attachment. It would be a significant insight if we could explain the experimental data based upon solely intrinsic properties of the graph such as node degree and local clustering coefficient of a node, with reference to how these relate to

fundamental properties such as entropy and constraints. In the next section we propose an extension, based upon the concept of constraints to the maximum degree of a node. This constraint is motivated from real world concerns in many networks. For example in communications networks the number of physical connections a node can maintain has a hard limit, and even in social networks building a network of friends is subject to constraints of time and physical space. In Section 4 we show how both constraints and non-linear preferential attachment could arise from a deeper, more fundamental, entropic model.

## 2 A Pure Constraint Based Model

A core assumption of the scale free model is that new nodes attach to other nodes with a probability that is determined only by the degree of the target node; no other factors affect  $\Pi_i$  and attachment is unconditional. In most networks though this is not a fully accurate assumption, as most nodes will have some inherent upper limit on their capability to establish connections. We can imagine a network comprised of nodes capable of maintaining a maximum of  $c$  connections, with  $c_i(t)$  being the point in time capacity of node  $i$  at time  $t$ . To simplify the treatment we assume the capacity of all nodes is equal across the network. In this case we could imagine modifying the probability of attachment to account for the nodes capacity as they accumulate connections, with a multiplicative factor to the preferential attachment probability  $\Pi_i$ . This assumption of uniform maximum capacity is an approximation that we justify by the simplicity of the theoretical analysis it permits. We seek to avoid introducing a family of free parameters, which would equate to a family of constraints, to preserve the theoretical elegance of the treatment. When we come to compare our *constrained* model to real world data it does require us to make reasonable estimates for the effective average constraint. We assume that this acts as a scaling factor for the attachment probability, similarly to the fitness factor introduced in the Barabási-Bianconi model [6],[16], in essence acting like a conditioning of the probability of attachment with the probability the node can accept the connection. In the most general sense, we can write this as the ratio of the nodes capacity relative to the time varying, average capacity of an arbitrary node,  $\langle c(t) \rangle$  as:

$$\begin{aligned} \Pi_i^c &= \zeta_i \times \Pi_i, \text{ where } \zeta_i = \frac{(c - k_i(t))}{\langle c(t) \rangle} \\ &\text{and } \Pi_i = \frac{k_i(t)}{2mt} \end{aligned} \quad (7)$$

To calculate  $\langle c(t) \rangle$ , we observe that at any time  $t$  a given node  $i$  will have an expected value of capacity  $\langle c_i(t) \rangle = \langle c - k_i(t) \rangle$ . As we assume that  $c$  is a shared maximum capacity across all nodes this reduces to  $\langle c_i(t) \rangle = c - \langle k_i(t) \rangle$ , and we note that  $\langle k_i(t) \rangle$  is the expected value of a node's degree  $k_i = \langle k_i \rangle$ , which will be useful in section 3 when we will compare our constrained model against

real networks. We can also estimate the expected value of the capacity of a node, by assuming a base uniform distribution of attachments in the absence of preference. After  $n$  nodes have been added, we will have added  $nc$  capacity to the graph, and consumed  $2nm$  connections. In the simplest case for the average capacity of a node, after adding a large number of nodes  $n$ , we note that the average capacity must evolve to a constant as following:

$$\langle c(t) \rangle = \frac{nc - 2nm}{n} = c - 2m. \quad (8)$$

Unfortunately as written this attachment probability is not sufficient as  $\sum_i \Pi_i^c \neq 1$ . This can be demonstrated by expanding Equation (7) as follows:

$$\begin{aligned} \sum_i \Pi_i^c &= \frac{1}{(c - 2m)2mt} \sum_i (c - k_i(t))k_i(t), \\ &= \frac{1}{(c - 2m)2mt} \left\{ c2mt - \sum_i k_i(t)^2 \right\}. \end{aligned}$$

If we define  $\delta$  as

$$\delta = \frac{\sigma}{\sum_i k_i(t)} = \frac{\sigma}{2mt} \text{ where,} \quad (9)$$

$$\sigma = \sum_i k_i(t)^2 - \sum_i \langle k_i(t) \rangle^2 \quad (10)$$

the normalization sum becomes,

$$\sum_i \Pi_i^c = 1 - \frac{\delta}{c - 2m}.$$

In general  $\delta$  could be a function of time and degree, but as an approximation in our model we treat it as a constant of the system. We test that assumption in the simulations presented later in this section, which indicate that it is valid to assume that  $\delta$  eventually stabilizes to a constant as the network evolves. We run these simulations of network growth to mimic the parameters for a selection of the real network data we analyze. Investigation of models where  $\delta$  is a function of time (and potentially  $k_i$ ) is an current avenue of research, and the subject of future work. For our attachment probability to be a valid probability measure we need to establish that  $\frac{\delta}{(c-2m)} \geq 0$  and that  $\frac{\delta}{(c-2m)} \leq 1$ . In the first instance the numerator of Equation (9), as defined in Equation (10), is the variance of  $k_i$  across the graph, and so is strictly positive. Providing that  $c > 2m$ , we can safely assume  $\delta \geq 0$ .

Regarding the upper limit of  $\delta$ , we can appeal to Popviciu's inequality (see [22]) for a bounded distribution, with  $k_{max} = c$  and  $k_{min} = m$ . This states:

$$\begin{aligned} \sigma &\leq \frac{1}{4}(k_{max} - k_{min})^2 \leq \frac{1}{4}(c - m)^2, \\ &\Rightarrow \frac{\delta}{(c - 2m)} \leq \frac{(c - m)^2}{8(c - 2m)mt}. \end{aligned}$$

For times  $t > \frac{(c-m)^2}{8m(c-2m)}$ , we then conclude that as required  $\frac{\delta}{(c-2m)} \leq 1$ . With these limits established, we can modify the attachment probability by adding in  $\delta$  to produce a form for the attachment probability, which sums to unity at each time step across all nodes, below:

$$\begin{aligned} \Pi_i^c &= \zeta_i \times \Pi_i, \text{ where } \zeta_i = \frac{(c + \delta - k_i(t))}{c - 2m} \\ &\text{and } \Pi_i = \frac{k_i(t)}{2mt} \end{aligned} \quad (11)$$

For convenience, we can further simplify the expression for  $\zeta_i$ , as follows:

$$\begin{aligned} \zeta_i &= \alpha \left( 1 - \frac{k_i(t)}{(c + \delta)} \right), \\ \text{where } \alpha &= \frac{c + \delta}{c - 2m}, \text{ or equivalently } \alpha = \frac{c + \delta}{c - 2\langle k_i \rangle}. \end{aligned} \quad (12)$$

We can now write the complete probability of attachment as:

$$\Pi_i^c = \frac{\alpha k_i(t)(c + \delta - k_i(t))}{2m(c + \delta)t}. \quad (13)$$

For comparison with the Barabási-Albert model, using  $\alpha = \frac{c+\delta}{(c-2m)}$  from equation (8) we can rewrite  $\Pi_i^c$  as follows:

$$\Pi_i^c = k_i(t) \frac{(c+\delta-k_i(t))}{2mt} \approx k_i(t) \frac{1}{2mt}, \text{ for large } c.$$

This recovers the standard Barabási-Albert model in the case that the constraint  $c$  is infinite and therefore does not interfere with the dynamics of the network's evolution. Following the continuum approach, and dropping the explicit time dependency of  $k_i$  for clarity, we can substitute this into equation (2), to obtain

$$\frac{\partial k_i}{\partial t} = m \Pi_i^c = \frac{\alpha k_i(c + \delta - k_i)}{2(c + \delta)t} = \frac{\alpha k_i}{2t} - \frac{\alpha k_i^2}{2(c + \delta)t}, \quad (14)$$

with the fraction multiplied out for convenience later. This is directly solvable by separating as follows:

$$\begin{aligned} \frac{1}{\alpha} \int \frac{dk_i}{k_i(c + \delta - k_i)} &= \frac{1}{\alpha(c + \delta)} \int \left\{ \frac{1}{k_i} + \frac{1}{c - k_i} \right\} dk_i \\ &= \frac{1}{2(c + \delta)} \int \frac{dt}{t}, \end{aligned}$$

whose solution is:

$$\log \left( \frac{k_i}{c + \delta - k_i} \right) = \frac{\alpha}{2} \log(t) + \theta,$$

or in simplified form

$$k_i = (c + \delta) e^{\theta} \left( \frac{t^{\alpha/2}}{e^{\theta} t^{\alpha/2} + 1} \right).$$

Following the continuum method in [1] we apply the initial condition that  $k_i(t) = m$  at time  $t = t_i$ , to obtain:

$$k_i(t) = \left( \frac{\rho(c + \delta) \left( \frac{t}{t_i} \right)^{\alpha/2}}{1 + \rho \left( \frac{t}{t_i} \right)^{\alpha/2}} \right), \quad (15)$$

$$\text{with } \rho \text{ defined as, } \rho = \frac{m}{c + \delta - m}.$$

Again, we note that as  $c \rightarrow \infty$ ,  $\rho(c + \delta) \rightarrow m$ ,  $\alpha \rightarrow 1$ , and so equation (6) reduces to

$$k_i(t) = m \left( \frac{t}{t_i} \right)^{1/2},$$

the standard result from the continuum analysis of Barabási and Albert [1],[3]. We then note that the probability that a node has degree  $k_i(t) < k$  is:

$$\begin{aligned} P(k_i(t) < k) &= P \left( t_i > \frac{\rho^{2/\alpha} (c + \delta - k)^{2/\alpha} t}{k^{2/\alpha}} \right) \\ &= 1 - P \left( t_i \leq \frac{\rho^{2/\alpha} (c + \delta - k)^{2/\alpha} t}{k^{2/\alpha}} \right). \end{aligned}$$

Assuming uniform probability for the choice of node introduction time  $t_i$  of  $\frac{1}{(m_0 + t)}$  we arrive at the expression:

$$P(k_i(t) < t) = 1 - \frac{\rho^{2/\alpha} (c + \delta - k)^{2/\alpha} t}{k^{2/\alpha} (m_0 + t)}.$$

Although somewhat more complex than the expression in [1] it is nevertheless simple to compute the distribution equation  $P(k) = \frac{\partial(k_i(t) < k)}{\partial k}$  to obtain the main result of our constrained model:

$$P(k) = \frac{2(c + \delta) \rho^{2/\alpha} t}{\alpha(t + m_0)} \left( \frac{(c + \delta - k)^{\frac{2}{\alpha} - 1}}{k^{\frac{2}{\alpha} + 1}} \right). \quad (16)$$

In appendix A we examine the asymptotic behavior of Equation (16), which verifies that by careful manipulation the standard result of the scale free model  $\gamma = 3$ , is recovered in the limit  $c \rightarrow \infty$ . Further, this analysis also indicates that the dominant contribution to degree distribution for  $k \ll (c + \delta)$ , produces a scale free log linearity with power law exponent  $\gamma = \frac{2}{\alpha} + 1$ . This equivalence to a more straight forward power law, but with an exponent  $\gamma < 3$  for values of  $k \ll (c + \delta)$  indicates that the presence of a constraint influences the behavior of our model even for nodes early in their evolution. This is a significant result and we make use of it to compare the predictions of our theory against real network data and simulations in section 3.

The result in equation (16) has some interesting implications, as the presence of a finite capacity  $c$  alters the scale

factor for the distribution of the nodes, *whilst* preserving the essential aspects of scale free behavior. By way of example, the data for the IMDB movie actor database, as presented in Table 1, is plotted in Figure 1b, along with results from a simulation of our model. The movie actor database naturally produces a graph by assigning a vertex for each actor and connecting two vertices when the actors have acted in the same film. Figure 1b contains a theoretical plot of the distribution taken directly from equation (16), using  $\langle k \rangle = 127$ ,  $c = 900$  and with initial conditions of  $m_0 = 100$ , which we take from Table 1. For this plot we set  $\delta = 205$ , which we take directly from the simulation, which we discuss in the next paragraph. The unmodified scale free model would give a value of  $\gamma$  of exactly 3, but our modification has an initial value of  $\gamma = \frac{2}{\alpha} + 1$ , which increases as  $k \rightarrow c$  and reaches a limit when  $k = c$ . To calculate  $\gamma$  we can take  $c = 900$  from the dataset in Table 1 and  $k = 127.33$ , with the estimated value of  $\delta = 243$  (we average the ratio of  $\delta$  to  $c$ ), to yield  $\gamma = 2.35$ , versus the measured value of 2.3 in [1] and 2.43 from our simulation. By comparison, to the scale free model, our approach predicts the value of  $\gamma$  to 2.29%, compared to 30.4% for scale free, a significant improvement. In addition, there is no explanation in the scale free model for the degree of a node in the graph having a maximum value.

To further verify our model, and in particular the assumption that  $\delta$  can be effectively treated as a constant, simulations were run using the form of preferential attachment probability in equation (13), for a network sharing the same parameters of maximum degree and average degree as the IMDB network. We present those results in Figure 1a. The simulation was run for a selection of initial parameters to assess the evolution of  $\delta$ , and in each case the value quickly converges to a constant. Turning to the simulation of degree distribution, in Figure 1b the essential scale free nature of the network obtained is visible on the log scale graph, as is the goodness of fit and agreement between the simulation with a theoretical plot of  $P(k)$  using the same simulation parameters. Using the techniques described in [23], we can measure  $\gamma$ , and obtain a value of 2.40 versus a calculated value from equation (16) of 2.41, which is in close agreement.

We also ran simulations for the Patents Citation graph (Figure 1c) and the Web Provider network (Figure 1d), which both produce similarly good results to the IMDB network in terms of the closeness of fit between the simulated and theoretically obtained  $P(k)$ . We can conclude that the constrained model is a good representation of networks with a simple maximum degree constraint.

Motivated by this example and simulation, in the following section we extend our analysis to a range of more recent, publicly available, network data to investigate further the accuracy of our constrained model.

### 3 Analysis and Comparison of Constrained versus Preferential Attachment

#### 3.1 Data and Methods

In this section we present the analysis of an extensive collection of network datasets comprising virtual, transport, and communications networks. The bulk of this data is publicly available through the Stanford Large Datasets Collection [24] which comprises an excellent repository of large graphs. The Twitter follower data is provided by [25], and the rest of the datasets are reproduced from publications such as [1], the Internet Topology Zoo [26]. We have one proprietary graph built from the topology taken from a large commercial deployment of network infrastructure used to deliver a top 10 Internet portal service (see [8]). The produced graphs fall into the following categories:

- *Social Networks*. These include Twitter, Facebook, Pokec graphs of the relationships between users. Typically each user is a node and nodes have links if the users have some form of relationship with each other. For example in the case of Twitter this relationship derives from one user ‘following’ another.
- *Collaboration and Citation Networks*. These cover a wide range of publicly available data, including the Arxiv citation, Patent Citation and co-authorship graphs as examples. Graphs are constructed by creating a vertex for each unique user or paper and then connecting the vertices if they share authorship with another vertex or directly cite it.
- *Communications Networks*. These networks, such as the Internet Router, IT Zoo, Web Provider and Berkeley Stanford Web Graph are constructed by representing physical or virtual nodes by a vertex in the graph and communications links as edges connecting the vertices.
- *Biological Networks*. These networks use a graph to represent a biological process, for example the metabolism of the *E. coli* organism. Nodes in the graph represent a molecule or intermediate state in the process used by *E. coli* to release energy from its food sources, with edges connecting nodes where a reaction or transition occurs. Similar networks exist for other biological processes (e.g. for the genetic cause and effect in cancers and disease epidemic spreads).

Analysis of the data was undertaken using a program and graph datastore which is available from the authors on request. The source data was often very large (the Twitter data contains for example over 10 million edges), and extracting values for the max degree and  $\langle k \rangle$  is not necessarily evident. Some of the data had some extreme outliers in terms of node degree, and to avoid skewing the results, we estimated the constraint at the 99<sup>th</sup> percentile of  $k$  rather than the maximum value in the data. This is consistent with the methodology taken in the theoretical analysis, where we made an assumption of the node degree constraint being constant for all nodes. This is a simplification, but one with great benefit in the analytical treatment of the model. The elimination of outliers



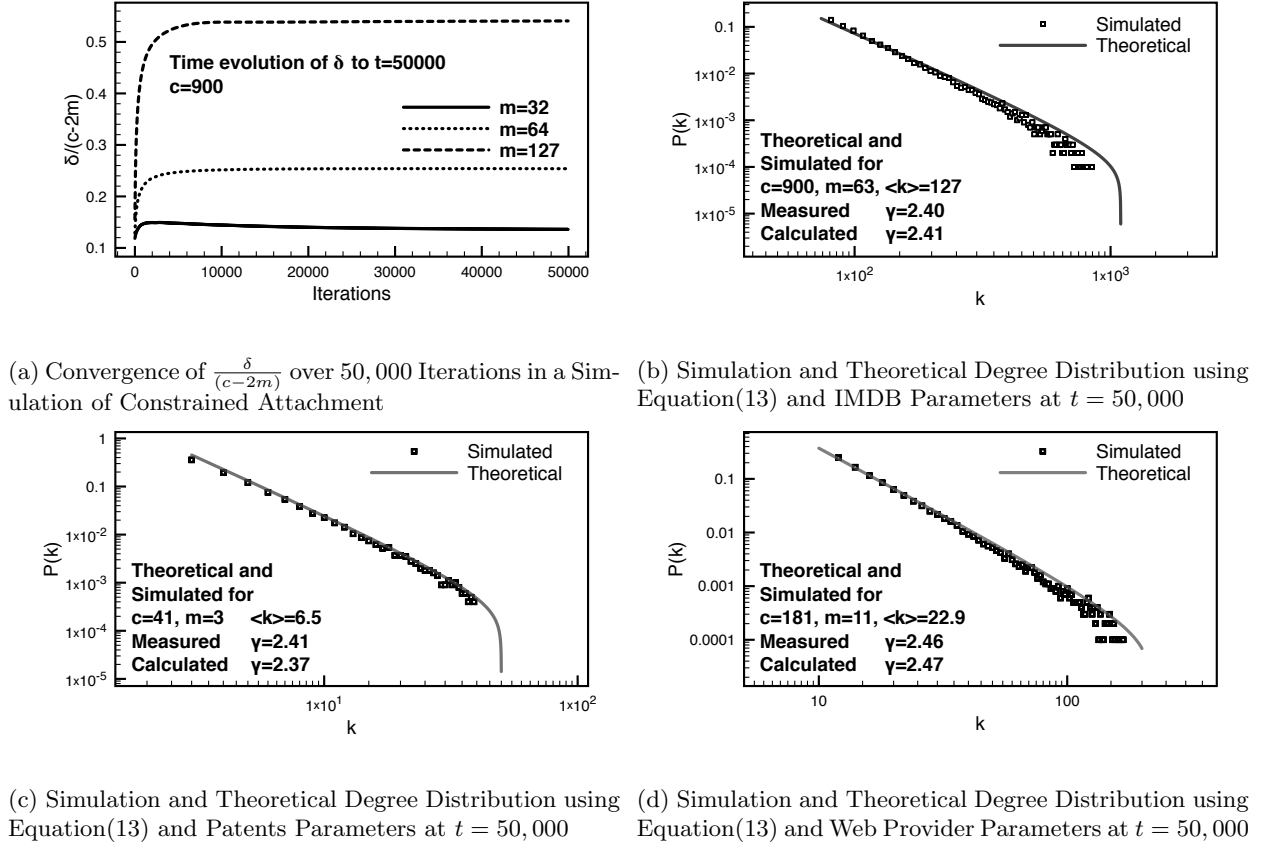
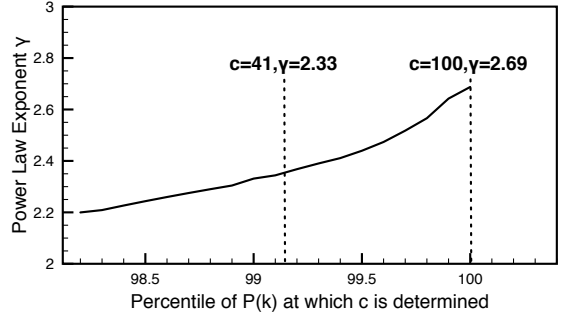


Fig. 1: Simulation Results for Constrained Attachment

at first sight may seem inconsistent with the assumption of a single constraint in the capacity of a node, but it is expected that the real world data will contain perhaps many different constraints, and that the average behavior of the graph will be most influenced by the effective maximum established at the 99<sup>th</sup> percentile. Further, the data above the 99<sup>th</sup> percentile in  $k$  is typically very sparse and may contain spurious data points, which this cut off eliminates. In Figure 2 we present the variation of the calculated value of  $\gamma$  with the choice of percentile at which to choose  $c$ . The range of calculated values as we move from the 98.2<sup>th</sup> to the 100<sup>th</sup> percentile is 2.20 to 2.69, a range of  $\pm 9\%$  either side of the chosen value of  $c = 41$ . We believe this further strengthens our choice of the 99<sup>th</sup> percentile as the appropriate cut off for measuring  $c$ .

For  $\langle k \rangle$  we require the expected value of the degree. This was calculated by computing the weighted mean, a discrete approximation of  $\langle k \rangle$ , which is truly only valid if  $k$  is a continuous variable. This is consistent with the approximation of continuity inherent in the continuum analysis approach.

To compare against the actual value of  $\gamma$ , power law exponent, we followed the techniques outlined in [23] to both assess the presence of a scale free distribution and obtain the value of  $\gamma$ . For the datasets we analyzed, which can be seen visually in Figures 3, 4 and 5, there is a considerable portion of the distribution which has a well defined straight line on the log/log plots, illustrating the intrinsic

Fig. 2: Variation of Calculated Values of  $\gamma$  with Choice of Percentile for  $c$  for the Patents Graph

power law distribution of node degree. We capture the measured values of these power law exponents in Table 2.

### 3.2 Analysis

In the summary Table 2 it is compelling to note that in all but a few cases the constrained model is more accurate in its predictions of  $\gamma$  than the standard scale free model. Indeed in the case of the Patent Citation, Internet Topology Zoo, Pokec, the real world network from a Web Provider, and a number of the citation networks and social networks, it comes very close to an exact prediction.

Table 2: Comparison of  $\gamma$  Predictions Between Preferential Attachment and Constraints Model

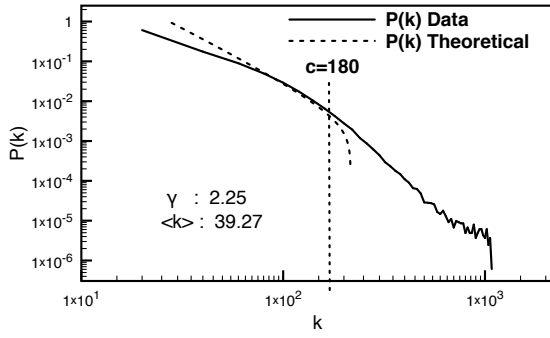
Source	$\langle k \rangle$	$c$	$\gamma$ Calculated	$\gamma$ Measured	$\Delta$ Constraints	$\Delta$ Scale Free
Patent Citation <sup>2</sup>	6.57	41	2.33	2.31	<b>0.75%</b>	29.66%
IT Zoo <sup>1</sup>	2.26	10	2.32	2.36	<b>1.63%</b>	27.19%
Internet Router <sup>7</sup>	2.57	30	2.44	2.48	<b>1.64%</b>	20.97%
Arxiv - Condensed Matter <sup>2</sup>	9.13	51	2.32	2.37	<b>2.18%</b>	26.39%
IMDB Movie Actors <sup>7</sup>	127.33	900	2.35	2.30	<b>2.29%</b>	30.43%
Pokec <sup>3</sup>	39.27	180	2.30	2.25	<b>2.29%</b>	33.34%
Airport Connections <sup>9</sup>	11.18	126	2.35	2.29	<b>2.82%</b>	31.19%
Arxiv - HepTh (Cit) <sup>2</sup>	26.75	165	2.34	2.44	<b>3.82%</b>	23.10%
Twitter (Circles) <sup>4</sup>	33.94	264	2.37	2.47	<b>3.90%</b>	21.43%
Arxiv - HepTh (Collab) <sup>2</sup>	22.05	285	2.37	2.51	<b>5.37%</b>	19.67%
Web Provider <sup>5</sup>	4.18	36	2.09	2.23	<b>6.33%</b>	34.48%
Co-authors, math <sup>7</sup>	3.90	400	2.69	2.5	<b>7.60%</b>	20.00%
Berkeley Stanford Web <sup>6</sup>	24.59	173	2.31	2.35	10.45%	27.74%
Metabolic, <i>E. coli</i> <sup>7</sup>	53.51	137	2.47	2.20	12.20%	36.36%
AS Skitter <sup>2</sup>	54.13	150	1.94	2.34	17.27%	28.14%
Facebook <sup>4</sup>	42.99	198	2.25	2.75	18.08%	9.23%
Arxiv - Astro Phys <sup>2</sup>	23.81	144	2.33	2.87	18.70%	4.61%
Co-authors, neuro <sup>7</sup>	11.54	400	2.53	2.1	20.42%	42.86%
Enron Email <sup>6</sup>	40.25	280	1.84	2.42	23.87%	24.02%
Twitter (Follower) <sup>8</sup>	8.63	90	1.56	2.39	34.77%	25.46%
PA Road Network <sup>6</sup>	5.41	9	1.71	2.69	36.27%	11.71%
Co-authors, SPIRES <sup>7</sup>	173.00	1100	2.69	1.2	124.17%	150.00%

Given that the motivation to investigate the constrained model originated from considerations of network design in communications networks, it is interesting to see that this has some strong applicability to non-physical networks. We also present the analysis both as a collection of log/log distribution graphs in Figures 3, 4 and 5 and also summarize the key prediction of  $\gamma$  against the standard value of 3.0 from preferential attachment in Table 2. In the log/log plots we overlay the value of  $c$  at 99<sup>th</sup> percentile, the average value of  $\gamma$  to this constraint and the expected value of the node degree  $\langle k \rangle$ . In each of Figures 3, 4 and 5, we also overlay the theoretical prediction for the distribution  $P(k)$  obtained by substituting the values of  $\gamma$  from Table 2 into Equation (16). The agreement between the predicted values of  $\gamma$  and the measured ones for our datasets is evident from these combined theoretical and experimental plots, at least for portions of the distribution. A consequence of the selection of  $c$  at the 99<sup>th</sup> percentile is that our theoretical curve displays a cut off earlier than the experimental data, which is to be expected.

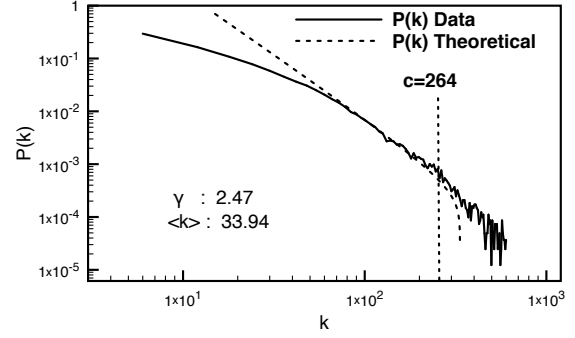
The striking feature of many of the degree distributions is the absence of strict linearity, contrary to the predictions of the standard scale free model, and also the marked increase in  $\gamma$  at high values of  $k$ , a key prediction of our constrained model and a necessary precursor to a hard constraint in the value of  $k$ . In the social network data we analyzed this is best illustrated in Figures 3a, 3c and 3b.

Similar behavior is also present in the citation network (perhaps the best example being Figure 4d), and again in the infrastructure graphs, particularly the Internet Topology Zoo (Figure 5a). It is interesting to speculate what the nature of the constraint is in the social networks, but this is perhaps explained by the effective limitations, no matter how small, on the amount of time people can feasibly spend on social networking platforms. Indeed in almost every conceivable network a constraint is a natural feature. Whether the node in the graph is a physical device, and individual engaged in an activity such as writing papers, or web site hyperlinks, there is a limitation to the connections a node can have. In some cases these are hard design limits such as ports on a network switch, in others it is simply the capacity of a human being, with a fixed lifespan, to blog, interact, star in a movie or engage in any other social activity. In every case our experimental data bears this out.

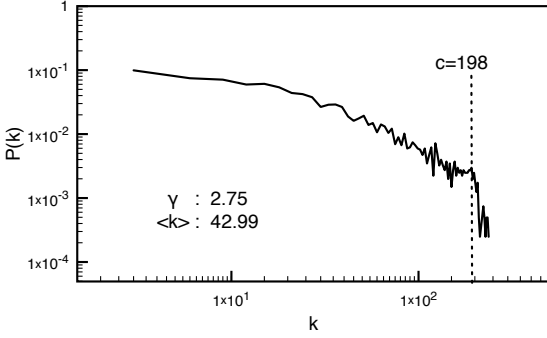
In the following Section 4 we point out how the two models may well be related to a fundamental dynamical principle that arises from thermodynamic considerations of network evolution. Critically this analysis derives the form of preferential attachment presented as an axiom in the scale free model.



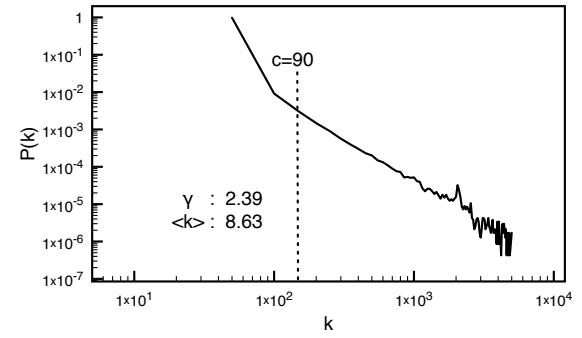
(a) Pokec - Slovakian Social network Friendship Graph, Theoretical and Experimental [28]



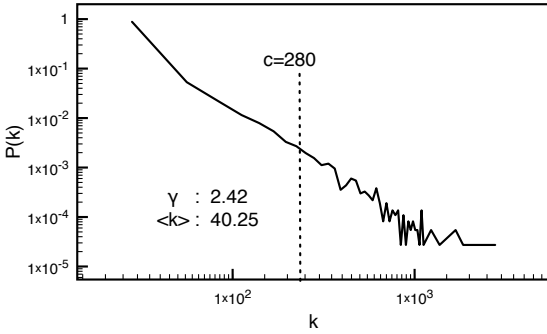
(b) Twitter Friendship Circles, Theoretical and Experimental [29]



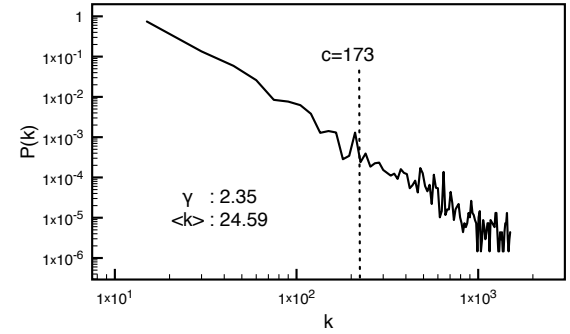
(c) Facebook Friendship Network [29]



(d) Twitter Follower Network [25]



(e) Enron Email Communication Network [30]



(f) Berkley Stanford Web Interconnection Network [30]

Fig. 3: Degree Distributions from Social Networking and Web Networks on a Logarithmic Scale

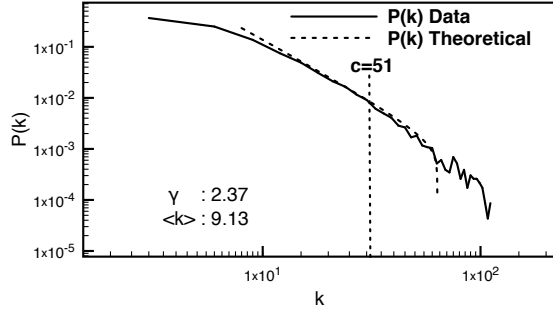
#### 4 Dynamical Evolution of Scale Freedom

In our treatment thus far we have followed the continuum model of Barabási-Albert with the addition of a constraint-based factor to the attachment probability. However, we can attack the problem from a more fundamental viewpoint. Essentially, we argue that the evolution of a graph satisfies the criteria for a treatment based upon considerations of entropy from a statistical mechanics perspective, in accordance with the 2<sup>nd</sup> law of thermodynamics. In any isolated physical system the entropy of the system will tend to a maximum unless energy is input to prevent that. For a classic treatment see [32]. In natural processes this tendency to increase entropy can be modeled as a macroscopic force on the system. This entropic force is responsible for both the elasticity of certain polymers and the biological process of osmosis. Indeed if thermodynamic

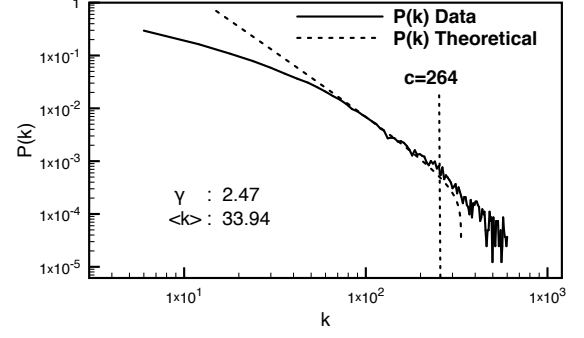
temperature is written as  $T$  and entropy  $S$ , one can state the entropic force  $F$  acting on a body when a process changes entropy as follows:

$$F = T \Delta S. \quad (17)$$

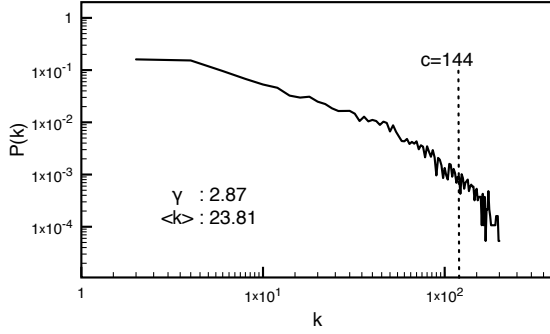
To begin our treatment of graph evolution from fundamental thermodynamic principles, it suffices to pose the problem in an appropriate manner. Consider an existing graph of  $m_0$  nodes and  $e_0$  edges in thermal equilibrium with an infinite supply of unattached nodes, each capable of connecting to  $m$  nodes in the event that it comes into contact with the existing graph. At every time-step we imagine that such an interaction occurs and the new node connects to  $m$  others. Our problem is to identify the probability of attachment for a node according to its degree  $k$ , and thus derive the degree distribution. More



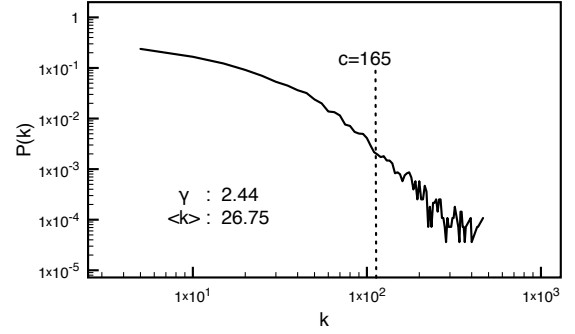
(a) Arxiv Condensed Matter Citation Network, Theoretical and Experimental [27]



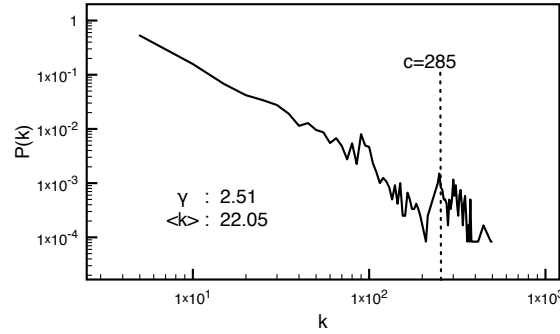
(b) Patent Citation Network, Theoretical and Experimental [27]



(c) Arxiv Astro-Physics Citation Network [27]



(d) Arxiv High Energy Physics Citation Network [27]



(e) Arxiv High Energy Physics Collaboration Network [27]

Fig. 4: Degree Distributions from Collaboration and Citation Networks on a Logarithmic Scale

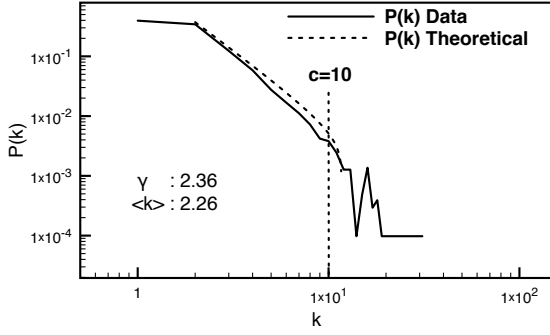
strictly, it is necessary to consider an ensemble of all possible graph configurations, *at every time step*, to enable statistical treatment of this process. This requirement to consider an ensemble of configurations is at first sight an added complication, but in fact is critical in permitting the analysis of the model. Whenever we consider a randomly selected node, for example in equation (18), it is important to recognize that we must average any interaction with the remaining graph over *all possible graphs* that can be constructed from the subgraph obtained by removing the randomly selected node and all edges connected to it. This ensemble average is further constrained by the total number of vertices and edges being unchanged after the removal of the random node. This requirement to average over all possible graph configurations at each time

step justifies the approximation we make to calculate, for example, the average clustering coefficient.

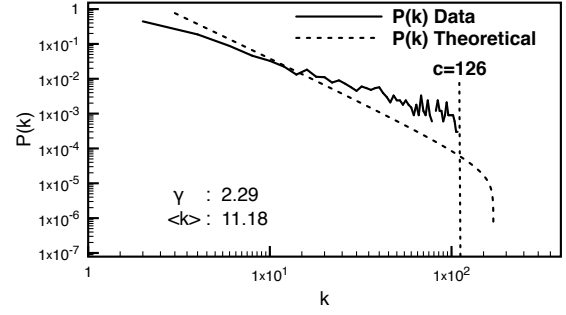
The probability of attachment to a random node must statistically and universally seek to maximize total entropy. Our model proposes that the probability of this random node acquiring new links is a result of the relative strength of the entropic force of attachment to the randomly chosen node versus any other node in the graph. Those nodes which exert the highest entropic force relative to the rest of the nodes in the network will gain the most links, and we write this mathematically as:

$$\Pi_i = \frac{F(v_i)}{\sum_{j \neq i} F(v_j)} \quad (18)$$

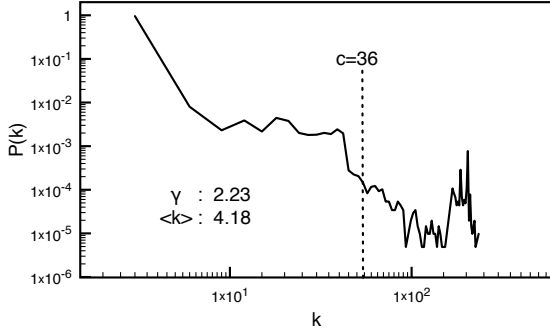
where  $F(v_i)$  is the entropic force of attraction to node  $i$ .



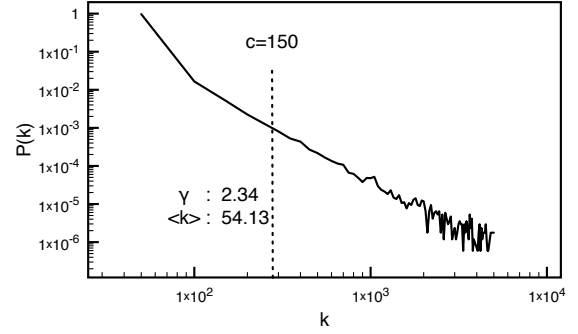
(a) Internet Topology Zoo Network, Theoretical and Experimental [26]



(b) Airport Flight Interconnection Network, Theoretical and Experimental [31]



(c) Web Provider Datacenter Network [8]



(d) Internet Autonomous Systems Network [27]

Fig. 5: Degree Distributions from Infrastructure and Communications Networks on a Logarithmic Scale

This expression governs the individual interaction that our randomly selected node has with a particular graph configuration, analogous to the elastic collision equations used to formulate the statistical treatment of ideal gases. In a similar way we cannot easily analytically formulate the dynamical equations of the graph from this equation as they are very large, and so to derive the degree evolution equations from this formulation we utilize statistical ensemble arguments. Considering all possible configurations of the graph  $G(V(t), E(t))$  at a fixed time  $t$ , the denominator of equation (18) is computed as an expectation value of the relative force of attaching to any other node, across all possible graphs at time  $t$  in the ensemble that our random node could be connected to. At a given time  $t$  in the evolution of the graph the numbers of vertices  $|V(t)|$  and edges  $|E(t)|$  are constant, but we do have to consider all possible graph configurations of that number of vertices and edges. This will ultimately change the average of the change in entropy that the node could make on connecting to any other node in the graph other than our randomly selected node  $v_i$ . In this way we collapse the denominator to the expected value of this entropy change, averaged across all possible connection points in all possible members of the ensemble. We write this as  $\mathbf{T} \times |V| \times \mathbb{E}(\Delta \mathbf{S})$ . As the graph becomes larger, we make the assumption that the value of  $|V| \times \mathbb{E}(\Delta \mathbf{S})$  is effectively constant, and factor this out. We base this assumption on the fact that most real world networks do indeed demonstrate some form of steep drop in the distribution of node degrees, so that the vast majority of nodes possess low degree (an impor-

tant claim of [4] and [1]). It seems reasonable to assume that with such a restricted degree sequence most nodes will contribute a similar amount to the change in entropy, and this expected value will stabilize to a constant. More complex analysis could admit a time varying value of this constant, as strictly both  $V$  and  $\mathbb{E}(\Delta \mathbf{S})$  may have complex time dependence, but for simplicity we assume:

$$\epsilon = \frac{1}{|V| \times \mathbb{E}(\Delta \mathbf{S})}.$$

With this assumption equation (18) simplifies and  $\mathbf{T}$  factors out to yield

$$\Pi_i = \epsilon \Delta \mathbf{S}_i. \quad (19)$$

In general  $\mathbf{S}_i$  is a function of potentially many variables  $x_i$ , but certainly depends upon  $k_i$  and time  $t$ . We can calculate  $\Delta \mathbf{S}_i$  as a total differential,  $\Delta \mathbf{S}_i(x_j) = \sum_{x_j} \frac{\partial \mathbf{S}_i}{\partial x_j} \Delta x_j$ , but we can assume for simplicity that  $t$  is fixed and the dependence is purely upon  $k_i$ . In this case  $\Delta \mathbf{S}_i = \frac{d\mathbf{S}_i}{dk_i} \times \Delta k_i$ , with, for a single time step,  $\Delta k_i = 2m$ . This gives us our expression for attachment probability:

$$\Pi_i = \epsilon 2m \frac{d\mathbf{S}_i(k_i)}{dk_i}. \quad (20)$$

To make use of equation (20) we require an expression for the entropy of a node in the graph. The subject of the entropy of a graph has a long history, originating in the work of Körner on the informational entropy of signals

described in [33] and [34]. Many approaches to calculating the entropy of a graph have been proposed, including the use of the eigenvalues of the adjacency matrix (see [35], and ensembles of networks with similar degree sequences (proposed in [36]). Unfortunately these concepts relate to the global value of entropy for a graph, and do not have utility when calculating the change in entropy as a new node connects.

A series of papers by Dehmer ([37],[38]) formalized the concept of the individual entropy of a node. In recent work [8] we built upon this formulation to define a local vertex measure (referred to in [8] as  $NVE'$ , and equivalent to our definition of  $\mathbf{S}_i$  here) in terms of its relative degree as:

$$\mathbf{S}_i(k_i, t) = \frac{1}{C_i^1} \times \frac{k_i}{2|E(t)|} \log \frac{2|E(t)|}{k_i}, \quad (21)$$

where  $C_i^1$  represents a modified clustering coefficient of the 1-hop neighborhood of the node  $v_i$ . Contrary to the more common point-deleted neighborhood clustering coefficient,  $C_i^1$  preserves the node in the calculation to measure similarity to the local perfect graph  $K_n$  of order  $n = k_i + 1$ . For convenience we give an explicit definition of the 1-hop neighborhood  $N_i^1$ :

$$N_i^1 = \{v \in V \mid d(v_i, v) \leq 1\} \cup \{v_i\},$$

and the related '1-edges'  $E_i^1$  as

$$E_i^1 = \{e_{jk} \in E \mid v_j \in N_i^1 \text{ and } v_k \in N_i^1\}.$$

We can then define the modified clustering coefficient to be

$$C_i^1 = \frac{2|E_i^1|}{k_i(k_i + 1)}. \quad (22)$$

At this point we can make use of the fact that we must consider all possible intermediate graph configurations to assume effective uniformity in the graph to calculate  $|E_i^1|$ , and assert that for a given node,  $|E_i^1| = \frac{k_i + 1}{|V|} \times |E(t)|$ . This then yields for the clustering coefficient the following expression:

$$C_i^1 = \frac{2|E(t)|}{k_i|V(t)|}. \quad (23)$$

Given that at every time-step we add one node to the graph, connecting to  $m$  other nodes we can write  $|V| = m_0 + t$ , and  $|E| = e_0 + mt$ . In general as the model evolves,  $t \gg m_0$  and similarly,  $mt \gg e_0$ , these simplify to  $|V| = t$  and  $|E| = mt$ . Substituting back in we obtain the following equation for vertex entropy at  $v_i$  at time  $t$  as:

$$\mathbf{S}_i(k_i, t) = \frac{k_i^2}{4m^2t} \log \left( \frac{2mt}{k_i} \right). \quad (24)$$

In the analysis undertaken by Tee *et al* in [8,9], this quantity was identified as sharing some of the properties of the structural entropy of the graph when summed across all vertices. In particular, the extremal behavior of the summed vertex entropy was proven to be minimized by

the perfect graph of order  $n$ ,  $K_n$ , and maximized by the star graph of order  $n$ ,  $S_n$ , for simply connected undirected graphs. From the perspective of dynamical evolution of networks, this is consistent with the approach in our analysis. The perfect graph  $K_n$  will tend towards a more node level disordered graph such as  $S_n$  as addition of nodes selects targets such as to increase the value of  $\mathbf{S}_i$  in Equation (24). From a purely statistical mechanics perspective one can consider each connected graph on  $n$  nodes and  $|E|$  edges as representing a micro-state. The perfect graph is achievable in precisely one unique configuration if edges are indistinguishable, whereas other configurations,  $S_n$  for example, can be achieved by selecting any one of the nodes as the hub vertex. In this way the result that increases in entropy tends to destroy cliques and regular ordered graphs is consistent. From this perspective we would expect dynamic processes to favor the attachment to nodes where the increase in  $\mathbf{S}_i$  is greatest. From here it is straightforward to follow through the continuum analysis as described in [1]. For the time evolution of  $k$  the following equation, is obtained:

$$\frac{dk_i}{dt} = 2m\Pi_i = -\epsilon \frac{k_i}{t} \left\{ \frac{1}{2} + \log \left( \frac{k_i}{2mt} \right) \right\}. \quad (25)$$

Although at first sight this nonlinear ODE appears intractable, in fact an analytic solution is available. Making the change of variables  $y = \log k$  and  $x = \log t$ , so that  $\frac{dy}{dx} = \frac{t}{k} \frac{dk}{dt}$ , we see that (25) becomes

$$\frac{dy}{dx} = -\epsilon \left[ \frac{1}{2} + y - \log(2m) - x \right]$$

This is now a linear ODE which can be solved by standard methods. Applying the initial condition  $k_i(t_i) = m$  the solution is found to be most conveniently expressed in the form

$$\log k_i(t) = \log(2mt) - \frac{1}{2} - \frac{1}{\epsilon} + \left[ \frac{1}{2} + \frac{1}{\epsilon} - \log(2t_i) \right] \left( \frac{t_i}{t} \right)^\epsilon \quad (26)$$

For values of  $\epsilon < 1$  the behavior of  $k_i(t)$  is similar to the Barabasi–Albert model: degrees increase monotonically but at an ever decreasing rate. An analytic form for the degree distribution, analogous to (3) does not seem straightforward to derive.

Figure 6 compares numerically computed degree distributions from the model (26) (shown in figure 6a) and the Barabasi–Albert model, shown in figure 6b. In each case a new node was added to the network every 0.5 time units, setting  $m = 5$  and growing the degrees of existing nodes according to (26) or (3) respectively. Degree distributions are plotted for fixed end times  $t_{end}$ , taking the values  $3 \times 10^2$ ,  $10^3$ ,  $3 \times 10^3$ ,  $10^4$ , and  $3 \times 10^4$ . The degree distributions for the entropy-based model do not clearly follow any power law behaviour, at least in the regime explored here, while the Barabasi–Albert model quickly assumes a form very close to a power-law degree distribution with exponent  $\gamma = 3$  as we expect.

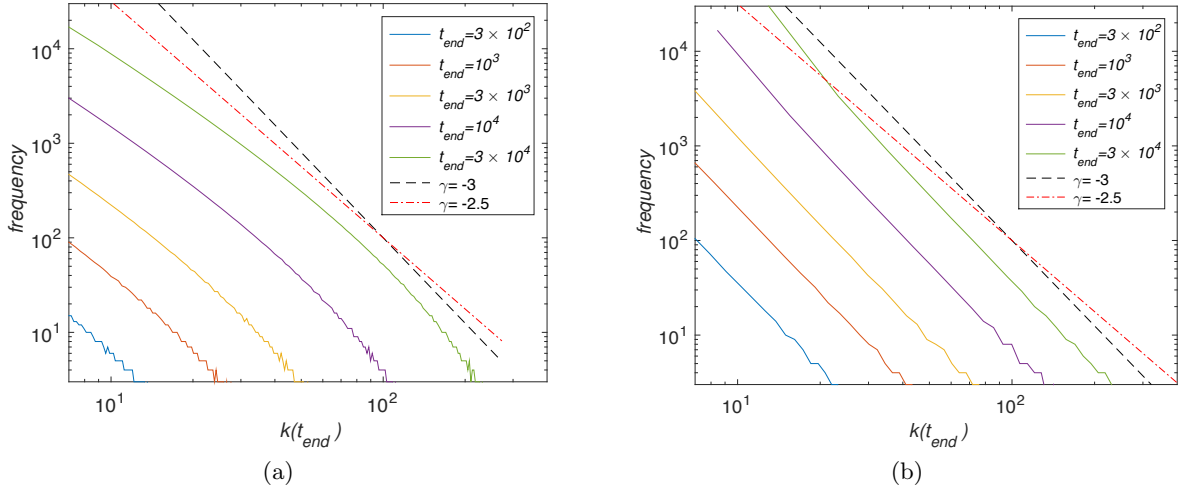


Fig. 6: Degree distributions for growing networks at fixed end times  $t_{\text{end}} = 3 \times 10^2, 10^3, 3 \times 10^3, 10^4$ , and  $3 \times 10^4$ . (a) Entropy-based model. (b) Barabasi-Albert model. In both cases a new node is introduced every  $\Delta t = 0.5$  and the node degrees evolve according to (26) or (3), in (a) and (b) respectively. Parameter values:  $m = 5$ ,  $\epsilon = 0.1$ . For illustration we have plotted the power law distribution lines at  $\gamma = -3$  and  $-2.5$ .

While any systematic analysis of (26) seems difficult, for large enough networks we might expect that this model is comparable to the classes of sub-linear preferential attachment models studied rigorously by Dereich & Mörters [39,40]. These authors prove that preferential attachment rules based on concave functions of node degree will asymptotically result in degree distributions with exponent  $\gamma = 3$ . This suggests that the long time dynamics of the entropy-based model might also show this behavior, but at intermediate times the more complex distributions illustrated in figure 6(a) might well be more typical.

## 5 Conclusion and Future Directions

In Section 2 we introduced a modification to the preferential attachment model to account for the maximum connections a node may have in a network. From the mathematical analysis we were able to predict both the value of the power law exponent  $\gamma$  and the presence of a hard limit on the degree distribution. In Section 3 we applied the analysis to an extensive range of social, citation and physical infrastructure graphs, and found that the constraint model's values for  $\gamma$  more accurately fitted the data. In addition, the constrained model implicitly contains a hard limit in the node degree, and the data analyzed had degree distributions with far fewer nodes of extremely large  $k$  than a pure power law would predict. This is an important result because the value is arrived at as a natural consequence of the presence of constraints on the maximum node degree, rather than by introducing a distribution of additional parameters such as in the fitness model. Fitness is a valuable concept, and indeed in further work it is intended to investigate the role of a top constraint in a model extended to include the concept of fitness, or indeed generalized in a similar way to the NGF models. In particular the analogy with Bose-Einstein statistical mechanics is interesting, and opens up many applications of

network science in more general theoretical physics, but the method outlined in this paper captures the essential features of real degree distributions without requiring the concept of fitness.

Motivated by the interesting results when applying concepts from statistical mechanics, and the results for vertex entropy arrived at in [8], we also set out to see if scale free models could be arrived at from pure thermodynamic principles of entropic force. In Section 4 we were able to obtain, from first principles, an evolution equation for the degree of a random node, which although soluble analytically, presents challenges when deriving the degree distributions according to the continuum analysis. The Taylor series for  $\log(x)$  converges only for values of  $x$  in the range  $0 < x \leq 2$ , but as  $k \leq 2mt$ , and, both terms are always strictly positive, we can safely expand the log term in equation (25). The validity of this expansion is not valid for  $k \ll 2mt$  as the series for  $\log(x)$  converges very slowly as  $x \rightarrow 0$ . However at early times after the introduction of the node into the graph,  $\frac{k}{2mt}$  will be closer to 1 and we can expand the log to yield:

$$\log\left(\frac{k}{2mt}\right) \approx \frac{k}{2mt} - 1 + \text{higher order terms.}$$

For the same period of time this expression is valid we can see that the leading terms in this expansion contribute to the ODE time evolution of  $k$  the following:

$$\frac{dk}{dt} \approx \frac{\epsilon k}{2t} - \frac{\epsilon k^2}{2mt} + \text{higher order terms.} \quad (27)$$

What can be asserted is that for a period of time after a node is introduced into the network its behavior will be governed by the first terms in this expansion, with much more complex behavior as the network evolves. This is illustrated nicely in Figure 6 obtained from our numerical simulations. These first two terms in the expansion are identical in form to the evolution of  $k$  with time in

the Barabási-Albert model, *and also* a correction identical in form to our constrained model. This would indicate that for small  $t$  the behavior of the entropic model should closely resemble scale free, with a correction for constraints. As  $t$  increases the model will become more complex.

The model introduces  $\epsilon$  as a free parameter, and it is a legitimate question to ask what the correct value of this should be. In the numerical simulations we chose, for illustrative purposes,  $\epsilon = 0.1$ . The choice of  $\epsilon$  will have a profound affect on the family of graphs that can emerge from the initial conditions and in particular the slope of the power law degree distribution obtained. For example, values of  $\epsilon > 1$  will tend to generate power laws with  $\gamma < 3$ , and conversely  $\epsilon < 1$  will produce  $\gamma > 3$ , at least in the regime where the first term of equation(27) dominates. Given that the origin of the parameter is in the relative entropic force of the graph compared to a randomly picked node of degree  $k$ , one could speculate that its value measures the relative affect of an additional link on the bulk of the graph to increase entropy compared to an individual node of varying degree. High values of  $\epsilon$  perhaps indicate relatively more homogeneous graphs than low values, indicating that degree distributions drop off more slowly the more ordered a graph's initial state. In future work we intend to investigate the dependency of graph evolution on  $\epsilon$  in more detail, and whether the more complex evolution behavior of our dynamic model has utility in revealing more detail on the internal structure of dynamically evolving graphs.

We believe that there is a deep connection between vertex entropy and the evolution of networks. An attractive feature of our model is that it predicts scale free and more complex network evolution behavior from a first principles argument without appeal to any heuristics, node by node parameters, or indeed a stated but not justified property of nodes to seek out other high degree nodes with which to preferentially attach. Instead we argue from the safety of the second law of thermodynamics to a model which reproduces the essential features of scale freedom, and also the constrained model which we demonstrated provides a better fit to the experimental data. It is possible that higher terms in the expansion of equation (25) could yield insight into the detailed evolution of networks, and provide powerful analytical tools to for example determine the age of a network. Nevertheless, it is attractive to speculate that scale freedom, and similar models, may be a manifestation of the second law of thermodynamics as applied to graph evolution.

Beyond investigating the entropic model, there are many potential enhancements to the constrained model. In further work we intend to conduct analysis of more network datasets and also investigate corrections to the constrained model to improve our estimate of  $(c - 2m)$  or  $(c - \langle k \rangle)$  for the average occupancy of a node, by iterating the resultant distribution in equation (16) to calculate  $\langle k \rangle$  as  $\langle k \rangle = \int_{-\infty}^{+\infty} kP(k)dk$ .

## Author contribution statement

P.T., I.J.W. and G.P. developed the initial model and wrote the initial draft of the paper. I.Z.K. and P.T. refined the exposition of the constrained attachment and continuum models and resolved some technical issues. J.H.P.D. provided subsequent analysis and the results shown in Figure 6. All authors contributed to editing and production of the final manuscript.

## A - Derivation of $\gamma$ in Constrained Attachment

We recall from the main body of Section 2 our expression for  $P(k)$  in Equation (16):

$$P(k) = \frac{2(c+\delta)\rho^{2/\alpha}t}{\alpha(t+m_0)} \left( \frac{(c+\delta-k)^{\frac{2}{\alpha}-1}}{k^{\frac{2}{\alpha}+1}} \right) \sim \frac{1}{k^\gamma}.$$

We can simplify this by collapsing the uninteresting details as follows:

$$P(k) = \frac{A(B-k)^{\frac{2}{\alpha}-1}}{k^{\frac{2}{\alpha}+1}}, \text{ where} \quad (28)$$

$$A = \frac{2(c+\delta)\rho^{2/\alpha}t}{\alpha(t+m_0)}, \text{ and } B = (c+\delta)$$

Now, as  $a^b = \exp\{b \log(a)\}$ , we can write  $(B-k)^{\frac{2}{\alpha}-1} = \exp\{(\frac{2}{\alpha}-1) \log(B-k)\}$ . Substituting back into Equation (28), and taking the logarithm of both sides, we obtain:

$$\log(P(k)) = \log A + \left(\frac{2}{\alpha}-1\right) \log(B-k) - \left(\frac{2}{\alpha}+1\right) \log(k).$$

We can further simplify by noting that  $\log(B-k) = \log\{B(1-\frac{k}{B})\} = \log B + \log\left(1-\frac{k}{B}\right)$ . We note that if  $k \ll B$ , either by taking small values of  $k$  or allowing  $c \rightarrow \infty$ , then  $\frac{k}{B} \rightarrow 0$ , so that  $\log(B-k) = \log B + \log(1+0) = \log B$ . Bringing this altogether we have:

$$\log(P(k)) = \log A + \left(\frac{2}{\alpha}-1\right) \log B - \left(\frac{2}{\alpha}+1\right) \log(k).$$

Taking the exponential of both sides we end with the main result:

$$P(k) = \frac{2(c+\delta)\rho^{2/\alpha}t}{\alpha(t+m_0)} \times \frac{(c+\delta)^{(\frac{2}{\alpha}-1)}}{k^{(\frac{2}{\alpha}+1)}}, \quad \text{which is of the form,} \quad (29)$$

$$P(k) \propto \frac{1}{k^{(\frac{2}{\alpha}+1)}}.$$

In Equation (29), we arrive at the familiar form of a scale free distribution with  $\gamma = 2/\alpha + 1$ . It is interesting to note



that, as  $c > 2m$ , by definition,  $\alpha \geq 1$  with equality in the limit that  $c \rightarrow \infty$ . This yields a range for the power law exponent  $\gamma$  as  $1 \leq \gamma \leq 3$ , with the familiar result of  $\gamma = 3$  recovered in the case of the constraint being infinite, and therefore unimportant to the dynamics of the network growth.

We can also examine Equation (28) in the asymptotic limit of  $c \rightarrow \infty$ . We recall that  $\rho = \frac{m}{c+\delta-m}$ , and that  $\alpha = \frac{c+\delta}{c-2m}$ . At the limit  $c \rightarrow \infty$ ,  $\alpha = 1$ , which reduces Equation (28) to:

$$P(k) \approx \frac{2c(\frac{m}{c})^2 t}{(t+m_0)} \times \left\{ \frac{c}{k^3} - \frac{1}{k^2} \right\},$$

which multiplying out and allowing  $c \rightarrow \infty$ , gives (30)

$$P(k) \approx \frac{2m^2 t}{(t+m_0)} \times \frac{1}{k^3}.$$

As expected, this is precisely the form of the degree distribution in the standard preferential attachment model, which emerges as the constraint becomes infinite, and therefore unimportant in the dynamical growth of the network.

## References

1. R. Albert, A.L. Barabási, Review of Modern Physics **74** (2002)
2. B. Bollobás, *Random Graphs*, 2nd edn. (Cambridge University Press, 2001), ISBN 9780521797221, <http://dx.doi.org/10.1017/CB09780521797221>
3. Á. Albert, H. Jeong, **272**, 173 (1999), 9907068v1
4. D. Watts, S. Strogatz, Nature **393**, 440 (1998), 0803.0939v1
5. S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Physical Review Letters **85**, 4633 (2000), 0004434
6. G. Bianconi, A.L. Barabási, Physical Review Letters **86**, 5632 (2001), 0011224
7. P.L. Krapivsky, S. Redner, F. Leyvraz, Physical Review Letters **85**, 4629 (2000), 0005139
8. P. Tee, G. Parisi, I. Wakeman, *Towards an Approximate Graph Entropy Measure for Identifying Incidents in Network Event Data*, in *IEEE/IFIP Network Operations and Management Symposium, NOMS* (Istanbul, Turkey, 2016), pp. 1049–1054, ISBN 9781509002238
9. P. Tee, G. Parisi, I. Wakeman, IEEE Transactions on Network and Service Management **PP**, 1 (2017)
10. M. Faloutsos, P. Faloutsos, C. Faloutsos, In SIGCOMM pp. 251–262 (1999)
11. R. Albert, H. Jeong, A. Barabasi, Nature **406**, 378 (2000), 0008064
12. P.R. Guimarães, M.A.M. De Aguiar, J. Bascompte, P. Jordano, S.F.D. Reis, Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **71**, 3 (2005)
13. N. Berger, C. Borgs, J.T. Chayes, A. Saberi, Annals of Probability **42**, 1 (2014), 1401.2792
14. B.W. Herr, W. Ke, E. Hardy, K. Borner, Proceedings of the International Conference on Information Visualisation **2007**, 465 (2007)
15. R.M. D’Souza, C. Borgs, J.T. Chayes, N. Berger, R.D. Kleinberg, Proceedings of the National Academy of Sciences of the United States of America **104**, 6112 (2007)
16. G.B.a.L. Barabási, Europhysics Letters **54**, 13 (2000), 0011029
17. P. Moriano, J. Finke, Physical Review E - Statistical, Nonlinear, and Soft Matter Physics pp. 1090–1095 (2013)
18. G. Su, X. Zhang, Y. Zhang, EPL (Europhysics Letters) **100**, 38003 (2012), 1103.3196
19. J. Park, M.E.J. Newman, Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **70**, 1 (2004), 0405566
20. G. Bianconi, C. Rahmede, Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **93**, 1 (2016), 1511.04539
21. O.T. Courtney, G. Bianconi, Phys. Rev. E **95**, 062301 (2017)
22. R. Sharma, R. Bhandari, Communications in Statistics - Theory and Methods **43**, 4503 (2014)
23. A. Clauset, C. Rohilla Shalizi, M.E. J Newman, SIAM Review **51**, 661 (2009), arXiv:0706.1062v2
24. J. Leskovec, A. Krevl, *SNAP Datasets: Stanford Large Network Dataset Collection*, url{<http://snap.stanford.edu/data>} (2014)
25. M. Cha, H. Haddai, F. Benevenuto, K.P. Gummadi, International AAAI Conference on Weblogs and Social Media pp. 10–17 (2010)
26. S. Knight, H.X. Nguyen, N. Falkner, R. Bowden, M. Roughan, IEEE Journal on Selected Areas in Communications **29**, 1765 (2011)
27. J. Leskovec, J. Kleinberg, C. Faloutsos, ACM Transactions on Knowledge Discovery from Data **1**, 1 (2007), 0603229v3
28. L. Takac, M. Zabovsky, International Scientific Conference & International Workshop pp. 1–6 (2012)
29. J. Leskovec, J. McAuley, Advances in neural information processing ... pp. 1–9 (2012), 1210.8182
30. J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, Internet Mathematics **6**, 29 (2011), 0810.1355
31. J. Patokallio, *Open Flights*, <http://openflights.org/about> (2016)
32. E. Schrödinger, *Statistical Thermodynamics* (Courier Corporation, 1989), ISBN 0486661016, 9780486661018
33. J. Körner, *FredmanKomlós bounds and information theory* (1986)
34. G. Simonyi, Combinatorial Optimization **20**, 399 (1995)
35. F. Passerini, S. Severini, ArXiv e-prints p. 5 (2008), 0812.2597
36. G. Bianconi, EPL (Europhysics Letters) **81**, 28005 (2008), 0708.0153
37. M. Dehmer, Applied Mathematics and Computation **201**, 82 (2008)
38. M. Dehmer, A. Mowshowitz, Information Sciences **181**, 57 (2011)
39. S. Dereich, P. Mörters, Electronic Journal of Probability **14**, 1222 (2009)
40. S. Dereich, P. Mörters, Jahresber. Dtsch. Math.-Ver. **113**, 21 (2011)